THE PERCEPTION OF GOOD AND BAD NATURAL SCENE CATEGORY EXEMPLARS

BY

EAMON SEAN CADDIGAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Associate Professor Diane M. Beck, Chair
Assistant Professor Fei-Fei Li, Stanford University
Associate Professor Alejandro Lleras
Professor Daniel J. Simons
Professor John E. Hummel

Abstract

Images of natural scenes are easily categorized by human observers. Recent work has shown that “good” images, or those that are more representative of their category, are more easily categorized than “bad” ones. The present research investigates a novel hypothesis: that “good” images of scenes are more easily perceived than bad images. Participants performed a two-alternative forced choice task in which they indicated whether an image was an intact or phase-scrambled scene photograph. In this task, observers were able to “see” good images better than bad scenes, more accurately detecting their brief presentations. This effect is not influenced by prior knowledge about the categories used in the experiment. Scene inversion is also shown to have a similar effect on the intact/scrambled discrimination effect, but it does not interact with category representativeness, indicating that the advantage conferred by good exemplars is invariant to inversion. Finally, the good and bad images were analyzed using an objective estimate of image typicality, and this factor was also shown to predict observers' ability to detect the images. These results document a close relationship between natural scene categorization and detection, suggesting that rapid scene perception is strongly influenced by our experience with typical and representative environments.

This work is dedicated to the memory of Barry Chai.

Acknowledgements

I have been fortunate to work with many talented individuals whose support has made this project possible. Without the work and insights of Barry Chai and Ana Torralbo, this research would never have taken place. I must also thank my doctoral committee, Li Fei-Fei, Alejandro Lleras, Daniel J. Simons, John E. Hummel, and especially my advisor Diane M. Beck for many years of patient mentorship. Colin McHugh, Audrey Merz, and Cara Pawlowski provided crucial assistance in data collection. Finally, I want to extend my gratitude to many more friends, family members, and collaborators for their generous support over the past six years.

Table of Contents

Introduction

The speed with which human observers can extract meaningful information from natural scenes has impressed researchers for several decades. Early studies of the rapid apprehension of a scene's "gist" were performed in an attempt to understand what information observers acquire during the brief fixations that separate saccadic eye motions during normal viewing behavior. An influential series of studies investigated this issue using the sequential presentation of photographic slides (an RSVP stream) with the images' presentation times approximating typical fixation durations (113 ms – 333 ms). When observers were cued with a simple description of a target image (approximating the basic-level category of that image), they were accurate at detecting the presence of targets, even when the stimulus onset asynchrony (SOA) was at its most brief (Potter, 1976). Recent observations using forced-choice responses and backwards-masked presentations of single images has found similar levels of performance with SOAs below 30 ms (Walther et al., 2009; Greene & Oliva, 2009b). Given that perceptual processes can be conceptualized as an accumulation of information over time (Laming, 1968), these short display times imply either that observers can recognize a scene's gist after extracting very little information or this information can be accrued at a greater rate from natural scenes than other laboratory stimuli.

Good performance following a brief presentation does not necessarily require rapid perceptual processing. It simply means that the information necessary to determine gist can be extracted quickly, which tells us little about the time course of subsequent processes leading to recognition and response. However, not only is the gist information

extracted very quickly, but discriminative responses have also been show to be very fast. For instance, when accuracy is deemphasized, response times for scene categorization tasks can be extremely short while still supporting above-chance performance. When participants perform a go/no-go task while alternating between four scene categories, they can respond significantly above chance with response times below 300 ms (Rousselet, Joubert, & Fabre-Thorpe, 2005). Moreover, when participants are shown a unique scene in each hemifield, they are able to accurately execute a saccade to the side with a target image (a scene containing an animal) within 120 ms of stimulus onset (Kirchner & Thorpe, 2006). Similarly, the brain is known to distinguish between scene categories very quickly. The time course of scene-related decision processes has been measured through the analysis of event related potentials (ERPs) by identifying the time relative to stimulus onset at which the EEG can differentiate targets and non-targets. Response-related ERPs can differentiate target from non-target scenes as quickly as 150 ms after stimulus onset (Thorpe et al., 1996), suggesting that the scene categorization process can be carried out that quickly. Stimulus-driven differences in ERPs are observed even earlier, discriminating categories as quickly as 75-80 ms after stimulus onset (VanRullen & Thorpe, 2001).

In addition to being a very fast perceptual process, scene categorization can be achieved under conditions of limited visual attention. In one single vs. dual-task paradigm, participants performed an attentionally demanding search for a letter singleton at fixation and attempted to categorize a peripherally presented natural scene. Observers were able to perform equally well in the dual task condition as they had been in either of

the single task conditions, suggesting that the scene categorization task required none of the attentional resources that were engaged in the central visual search task (Li et al., 2002). Later investigators argued that some detection tasks can be performed on the basis of “unbound features”. Using the phenomenon of the attentional blink, it was demonstrated that attention was necessary for categorization tasks requiring greater depth of processing, such as “identifying” a target, rather than simply “detecting” one (Evans & Treisman, 2005). More specifically, an attentional blink was observed when observers were searching for animal targets and had to identify the specific type of animal (e.g., a gorilla), but no blink was present when participants only had to respond to the presence of any animal. However, it is currently unknown whether a putative feature binding process is actually necessary for any of the rapid scene categorization findings discussed here.

Natural scenes constitute an important class of stimuli for the study of human visual perception. We exist in a world of scenes, and our visual system is specially adapted to process them quickly with only a brief glimpse, and requires very little attentional resources to do so. The research described here will investigate natural scene perception by comparing scenes that observers are particularly adept at perceiving to those they are not.

Defining Natural Scenes

Throughout the literature, studies of “natural scenes” may utilize different types of stimuli. For instance, research on the speed of visual processing has generally used photographs of animals or vehicles in their typical environments and at a number of

spatial scales (e.g., Thorpe et al., 1996). Work investigating “scene space” (e.g., Oliva & Torralba, 2001) typically makes use of images depicting expansive environments along with closer views of natural textures. Many studies on the representation of spatial layout have actually used sparse arrangements of toys to approximate images of real-world environments (Epstein et al., 1999; Sanocki, 2003). For the sake of clarity, the following conventions will be used for the work presented here. A stimulus will be considered a “natural image” if it is a photograph that has undergone very little post-processing; natural images are snapshots of the world as they may be observed by a human viewer. Such stimuli are “natural” in the sense that they can be distinguished from the artificial stimuli used in many laboratory studies of human vision, such as geometric forms and gratings. A natural “scene” is a natural image that captures the environment such that the objects and surfaces it depicts are generally at the outer edge of “action space”, and extend into “vista space” (Cutting & Vishton, 1995). Natural scenes will typically be identified as an image of a specific type of environment (e.g., “forest” or “city”; Tversky & Hemenway, 1983), and not as an image of any one object (e.g., “tree” or “car”). There is an exception in the case of structures that define the environment and landscape features; although it can be debated whether buildings or mountains are objects or scenes, they will be used as scenes in the work here. The objects and environments depicted in natural scenes may themselves be natural or man-made; a “natural environment” is a scene that contains few if any man-made objects, with beaches and forests serving as common examples.

Scene Features

Which features from a natural scene come through quickly enough to support the rapid extraction of its gist? One set of results suggests that the information conveyed in low spatial frequencies is more important to rapid categorization than that carried by high spatial frequencies (Schyns & Oliva, 1994). When participants were shown images that had been processed with a low-pass or high-pass filter, they exhibited comparable levels of categorization performance at both short (30 ms) and long (150 ms) display durations. However, when presented with hybrid images composed of the low-spatial frequency component of an exemplar of one category and the high-spatial frequency component of an exemplar from a different category, categorization performance was qualitatively different at these display times: at short SOAs, participants categorized hybrids on the basis of their low spatial-frequency information, while at long SOAs, they responded to high spatial-frequency information. Preferential processing of low spatial frequencies may be an ideal strategy for natural scene recognition: the amplitude spectra of natural scenes follow a roughly $1/f$ distribution (Field, 1987). Natural scenes have the most power at low spatial frequencies, and therefore this region of frequency space carries more information.

The neuronal pathways originating in the magnocellular and parvocellular layers of the lateral geniculate nucleus (LGN) are thought to retain some degree of separation in neocortex. The M channel is responsible for the rapid transmission of low-spatial frequency, achromatic information to ventral and dorsal cortical areas (see Maunsell, Nealey & DePriest, 1990). Given that the removal of color information has been found to

have no impact on rapid categorization performance (Rousselet, Joubert, & Fabre-Thorpe, 2005), and that low spatial frequencies contain most of the information present in a scene, the M channel may be especially important in natural scene processing. Recent proposals suggest that the magnocellular pathway plays an important role in a model of top-down facilitation in object recognition (Kveraga, Boshyan, & Bar, 2007). In the model proposed by Bar (2003), low spatial frequency information from a visual stimulus is sent directly to orbito-frontal cortex, bypassing the hierarchical levels of processing associated with most bottom-up models of object recognition. Prefrontal activity generates a set of candidate representations that match this “blurry” image, which are integrated with the bottom-up activity via fast back-projections to areas associated with the traditional visual-processing pathway. This top-down information facilitates object recognition by biasing early visual processing in favor of these candidates.

If rapid gist extraction is heavily reliant on M channel neurons, it may be the case that top-down facilitation occurs as suggested by Bar's model. Recent work attempted to address the role of feedback connections in the perception of natural images using transcranial magnetic stimulation (TMS) to disrupt visual processing (Camprodon et al., 2009). In one experiment, participants were presented with natural images and asked to indicate whether each contained a bird or a large mammal. Images were presented for 14 ms, and after a variable SOA, a TMS pulse was delivered to the occipital pole. Accuracy was significantly lower at an SOA of 100 ms, which is consistent with previous reports of TMS-induced disruptions of early visual cortex (e.g., Kastner, Demmer, Ziemann 1998). This effect is thought to be due to the interruption of fast back-projections from later

cortical areas to V1, which are presumably necessary for conscious vision (Ro et al., 2003). However, accuracy was also found to be affected by pulses delivered at an SOA of 220 ms, which may be taken as evidence for the necessity of slower feed-back connections for conscious visual perception. This result is difficult to relate to fast gist extraction, as the timing of this second disruption is later than many “fast” scene perception results. Future work will be necessary to determine the precise role of this late volley of activity in the perception of scene gist.

In contrast to the view that low spatial frequencies are the key to scene gist recognition, another result suggests that the visual system is flexible in its use of spatial scale for scene perception (Oliva & Schyns, 1997). In a demonstrative experiment, separate groups of participants were exposed to hybrid images that contained scene information in either a high or low spatial frequency channel and structured noise in the other, sensitizing them to the channel that was taken from scenes. After this sensitization procedure, observers categorized hybrid images composed of low- and high-spatial frequencies of scenes drawn from different categories, and were found to make their decisions on the basis of whichever channel had been meaningful during the initial training. This result shows that viewers can use both low and high spatial frequencies to recognize the gist of scenes, and select whichever channel is the most informative in the context of the current experiment. The Spatial Envelope model of scene perception expands upon this, relying exclusively on spectral information across a larger range of frequencies (Oliva & Torralba, 2001). This influential model of scene representation was the first to describe a “scene space”, holding that images are represented along a number

of perceptual dimensions relating to the characteristics of environments, such as the properties “openness” and “ruggedness”. Computational analysis of scenes showed that these properties could be reliably predicted on the basis of the amplitude and orientation of the Fourier domain representation of the images. However, human observers were subsequently found to be incapable of recognizing the category of phase-scrambled images, which preserve this information (Loschky et al., 2007), and moreover, global property recognition is improved with coarsely-localized amplitude and orientation information.

The proposal that scenes are represented by global properties has received support from a number of behavioral findings. When viewers perform rapid categorization tasks, they are more likely to mistake the category of a target with another that has similar global properties than a wrong category with different global properties; for example, oceans are more likely to be confused with fields than waterfalls because the latter do not share the property of openness (Greene & Oliva, 2009a). In a two-alternative forced-choice classification task, participants are also able to classify the global properties of briefly presented and masked images at shorter display times than their basic level categories (Greene & Oliva, 2009b). Additionally, in contrast with traditional views of categorization which suggest that the basic level category of a stimulus becomes available prior to superordinate or subordinate levels (Rosch, 1975), viewers appear to make the distinction between man-made and natural environments prior to basic level categorization (Loschky & Larson, 2010). This distinction corresponds to the “naturalness” property proposed by the Spatial Envelope model.

The scene features extracted for use by the Spatial Envelope model are reminiscent of the receptive field properties of the cells in early visual cortex. Neurons here are known to be selective for spatial frequency (De Valois et al., 1982) and orientation within a restricted spatial extent (Hubel & Wiesel, 1968). The tuning of simple cells can be well explained by Gabor functions (Jones & Palmer, 1987), though it was for a long time an open question why the mammalian visual system would adopt this strategy to encode all visual information. A number of attempts to resolve this question examined linear decompositions of natural images to determine how a vision system could "best" represent the class of stimuli it most frequently encounters. A linear decomposition of a signal represents that signal through a weighted sum of basis functions. When principal components analysis is used to achieve this decomposition, images can be reliably reconstructed using only the several of the recovered basis functions, though the response properties of these functions bear little similarity to those of V1 cells (Hancock, Baddeley, & Smith, 1992). However, when the construction of the code must satisfy a "sparsity" constraint concurrently with the need to preserve information, a set of basis functions is found with properties similar to cortical simple cells (Olshausen & Field, 1996). This additional constraint ensures that only a few bases out of a larger set will be necessary to reconstruct any given image. This computational work indicates that the encoding scheme employed by V1 is therefore capable of representing natural images while engaging as few neurons as possible, which may in part explain why natural scenes are processed so rapidly.

What are the behavioral consequences of a sparse neural representation of natural

scenes? Local inhibitory interactions between cells generally result in mutual suppression of visual stimuli (Blakemore & Tobin, 1972). The Biased Competition model of attention states that the act of visually attending one stimulus instead of another competing item is achieved by "biasing" the activity in favor the item to be attended (Desimone & Duncan, 1995). If natural scenes are represented through sparse activity, then it could be the case that the features of a scene require less biasing to be resolved and represented, and therefore requires less attention. This is in line with results showing that some scene perception tasks have limited attentional resource requirements. However, the finding that attention is required for more difficult scene perception tasks is consistent with an account of the second volley of activity apparently uncovered by the TMS study of scenes (Camprodon et al., 2009) reflecting attention in the form of reentrant processing (Fahrenfort, Scholte, & Lamme , 2007).

Gist recognition without categorization

A traditional approach for determining the importance of a given feature in scene categorization is to present images in which that feature has been disrupted through an image processing procedure. The effect of this disruption is measured through changes in accuracy or response time in a scene categorization task. Although this approach has proven fruitful, there are disadvantages to using it alone. A potential confound is that the removal of a feature that is diagnostic for category distinction will necessarily lead to worse categorization, even though it could theoretically be the case that this feature is not key to the rapid formation of scene representations that are useful for other tasks, such as navigation or visual search. This issue may explain the apparently conflicting findings

regarding the relative importance of different spatial frequency channels in natural scenes: when the diagnosticity of a given channel is high for a categorization task (due to a sensitization procedure as used in Oliva & Schyns, 1997), disruption of this channel will lead to worse categorization. However, it may be the case that low spatial frequencies are always more important for instantiating scene representations, even if the representations aren't useful for making categorization decisions. The reliance on categorization tasks becomes more problematic when an experiment makes use of hybrid images that combine information from two different category exemplars. Although they have been used successfully in previous work (Schyns & Oliva, 1994), it is not clear what a “correct” categorization response would be to such a stimulus. It may be the case that participants, when asked to categorize such an image, are biased by demand characteristics toward choosing the response that they believe is sought by the experimenter.

An alternative to using categorization to measure the successful apprehension of a scene's “gist” is to measure how well an observer can “see” a briefly presented image, irrespective of its category. This can be measured through the use of a discrimination task in which viewers must indicate whether an image was meaningful or not. In the extreme case, this could be a 2AFC response to normal photographs versus white noise, though manipulation of the targets and choice of distractors are both potential sources of information; Experiments 1-3 use phase-scrambled images as foils (examples of phase-scrambled images are shown in Figure 1). Such a task has previously been described as a “detection” task in a series of studies that examined the relationship between such

discriminations and a more traditional categorization response (Grill-Spector & Kanwisher, 2005). The precise nature of the relationship between categorization and detection is currently a subject of debate. While early results suggested that both responses are essentially the same (Grill-Spector & Kanwisher, 2005), subsequent work has shown that discrimination responses are made faster than categorization responses for most choices of category (Mack & Palmeri, 2010). Preliminary results of a modeling approach suggest that observers make both responses on the basis of the same information accumulation process (therefore using the same neural representation). However, even if this is not the case, the ability to extract meaning from a briefly presented image seems as reasonable an operationalization of "gist" recognition as the identification of its basic-level category or a some global property such as "openness".

Good and bad category exemplars

Although natural scene categorization is fast and seemingly effortless, some exemplars may be described as being more representative of their respective categories; that is, some images will be seen as "good" exemplars of their category and others will not. Such a distinction may prove fruitful in understand gist processing. What is it about good exemplars that make them good? Are good exemplars actually processed better than bad exemplars? Recently, Torralbo and colleagues (in prep) found that scenes that are "good" exemplars of their category are categorized better than "bad" category exemplars. In this work, a collection of 4025 images from six basic level scene categories – beaches, city streets, forests, highways, mountains, and offices – were rated online by workers using Amazon Mechanical Turk (AMT). Images were rated on a five-point scale

according to how “representative” they were of their category, and these ratings were used to create nonoverlapping sets of 40 “good” scenes (which had high representativeness ratings) and 40 “bad” scenes (which had significantly lower ratings) for each category (examples are shown in Figure 1). These images were then used in a rapid six-alternative forced-choice categorization task, and observers had faster and more accurate responses to the good scenes.

This result is in line with “typicality” results described in object categorization (Rosch, Simpson, Miller, 1976); participants may simply be able to more quickly categorize good examples because they more readily evoke the concept of their category. However, the remarkable speeds with which the brain distinguishes among basic level natural scene categories – speeds which leave little if any time for feedback processes (Van Rullen & Thorpe, 2001; Walther et al., 2009) – raise another possibility. In particular, we propose that the visual system actually “sees” good category exemplars better. In other words, the very perception of the “good” exemplars is better than the “bad” ones. We test this novel hypothesis in Experiments 1-3 using a discrimination task in which observers are presented with intact scenes or phase-scrambled versions, and determining whether they are more accurate when the images are “good” exemplars rather than “bad” exemplars of a scene category.

We have reason to believe that such a perceptual advantage for good exemplars may exist based on the phenomenological reports of participants in the Torralbo et al. study. Image presentation durations in that study were adjusted for each participant so as to produce a 65% categorization accuracy, using an adaptive staircasing procedure. This

low staircasing threshold meant that participants' phenomenology was that very often they simply saw a "flash" with no coherent structure. However, participants also reported that occasionally some of the images "came in more clearly." We wondered whether those images that were seen "more clearly" were more likely to be the good category exemplars than bad category exemplars. Such a result would lend further credence to the idea that comparing good and bad category exemplars may be a useful tool for understanding how the human visual system is able to quickly process natural scenes, as the former may be processed more efficiently than the latter.

Good and bad natural scene category exemplars served as stimuli for this research. Experiments 1-3 employed a detection task to determine whether good scenes are "seen" better than bad scenes, and show an effect of representativeness on scene detection. Additionally, Experiments 1 and 2 compare the influence of scene representativeness in the case where participants know the list of categories used in the experiment – and are encouraged to consider the scenes as being exemplars of these categories – to the case where observers are naïve to the use of specific categories. Experiment 3 investigates the effect of scene inversion on a detection task and its interaction with representativeness; if categorization ability is closely related to detection accuracy, inversion, which has previously been shown to disrupt scene categorization accuracy (Walther et al., 2009), should also hurt detection performance. Finally, the images used in these studies are analyzed in an attempt to determine what aspects of the good scenes make them easier to detect than the bad scenes. Scenes that were rated as

being “representative” of their category by observers are shown to be more similar to other images from the same category, and this is shown to predict detection accuracy.

Experiment 1

Do participants "see" images that have been rated as good category exemplars better than images that have been rated bad category exemplars? Participants performed a two-alternative forced choice discrimination task, indicating whether a briefly presented image was intact or phase-scrambled (Sadr & Sinha, 2004). Viewers' ability to successfully discriminate an unaltered photograph from a phase-scrambled image would imply the detection of coherent local structure in that image (Loschky et al., 2007), while a failure to do so would indicate that an image of a natural scene was not perceived as such. In other words, sensitivity to the intact/scrambled distinction provides a measure of whether a coherent image was detected or not.

In an attempt to make the category of the images relevant to observers, we instructed participants to retain the list of categories used in the experiment and then, after each intact/scrambled discrimination response, rate the same image in relation to its category; in particular, they were to indicate how well the preceding image exemplified its category on a five-point scale.

Method

Participants.

18 participants from the University of Illinois took part in these experiments for course credit in an introductory psychology course. All had normal or corrected-to-normal vision and gave written informed consent according to the procedures of the University of Illinois Institutional Review Board.

Stimuli.

Full-color natural images were drawn from a set of 4025 images of beaches, city streets, forests, highways, mountains, and offices. Each image was rated to indicate how representative it was of its category by workers via the internet (Torralbo et al., in prep). Briefly, workers using Amazon Mechanical Turk rated each image on a scale from one ("Poor") to five ("Good"), or indicated that it did not belong to the specified category (see Torralbo et al., for more details). For each of the six categories, we selected 40 "good", 40 "medium", and 40 "bad" images based on their mean ratings (mean scores were 4.70, 3.99, and 2.88 respectively). Scenes were phase-scrambled by combining in the Fourier domain the amplitude of an intact scene with the phase from a random noise image, and taking the inverse Fast Fourier Transform of this hybrid image. Examples of intact and scrambled good and bad exemplars are shown in Figure 1. All images were presented at a resolution of 800 x 600 pixels and subtended approximately 30 degrees of visual angle.

Procedure.

Before beginning the task, participants were presented with a list of the categories used in the experiment and asked to use these categories when rating how well the images exemplify their category. After being instructed on the task, participants performed 25 blocks of 30 trials each. The first nine blocks were used for staircasing SOA and consisted of "medium" category exemplars drawn randomly with replacement. The stimulus onset asynchrony (SOA) between target image and mask was staircased to 70% accuracy individually for each participant using the QUEST algorithm (Watson &

Pelli, 1987). The SOA in the staircasing phase of the experiment began at 500 ms and was adjusted over the course of 270 trials to produce an accuracy rate of approximately 70%. There was no interstimulus interval between image and mask, thus adjusting the SOA amounted to adjusting the duration of the target image. SOA for the remaining 16 "testing" blocks was fixed at the mean of the probability density function obtained during staircasing (34 ms – 78 ms, mean across the experiments = 49 ms). The testing blocks consisted of "good" and "bad" category exemplars drawn randomly without replacement from each of the 6 categories. Each trial proceeded as follows: a fixation cross was presented at the center of the screen for 500 ms, followed by the presentation of the target intact or phase-scrambled image (with a fixation cross superimposed) at the SOA determined during staircasing. Immediately following the image, a perceptual mask was presented for 500 ms. Participants then indicated whether the image was intact or phase-scrambled (each condition accounting for 50% of the trials) by pressing one of two keys on a computer keyboard. No feedback was given.

During the testing phase of the experiment, participants performed an additional task after making each intact/scrambled response. Participants were asked to rate how well each image exemplified its category by pressing a number between one ("poor example") and five ("very good example"). Instructions given at the beginning of the experiment described this task, and participants were told to covertly categorize each image in order to make the judgment.

Results and Discussion

Overall, participants were 86% accurate on the intact/scrambled distinction and

needed only an average image duration of 53 ms to achieve that accuracy. Sensitivity for intact images was measured by calculating d' for subjects' intact/scrambled responses, with images correctly identified as intact classified as hits and those scrambled images falsely labeled as intact classified as false alarms. In keeping with our predictions, we observed a significant difference between d' for good and bad images (2.30 vs. 2.10; $t(17) = 2.95$, $p < 0.01$), such that participants were better able to discriminate intact from scrambled images if they were good category exemplars than if they were bad. We stress that in order to make this distinction participants need only detect some coherent structure in the image to be able to rule out that it was phase-scrambled. Mistakes occur because presentations are so brief observers often experience just a flash. Importantly, however, they were less likely to experience this incoherent flash if the image was a good exemplar of its category. Moreover, this pattern of results was apparent for all categories except for forests, which also had the lowest d' scores over all (see Figure 2a) presumably because the relatively uniform textures and preponderance of green hues of forest images made them harder to distinguish from their phase-scrambled versions. This difficulty may even have been exacerbated for good exemplars over bad because good exemplars of forests were more likely to include densely packed and fairly uniform views of multiple trees from within forest whereas bad exemplars had a higher occurrence of a focal tree or some structure that rendered it less uniform.

The image ratings used to determine good and bad category exemplars were obtained with different participants and under different viewing conditions than those used in the current study. It is therefore possible that this distinction would be lost due to

the short presentation times and perceptual masking present in this experiment. However, we again observed a higher average good/bad rating for intact good images than that for bad (4.07 vs. 3.70; $t(17) = 5.79$, $p < 0.01$; Figure 3a) in the rating task performed in the present experiment. This effect was not driven by the fact that participants failed to see more of the bad exemplars as intact, as the difference remained significant when only trials with correct "intact" responses were considered (4.13 vs. 3.85; $t(17) = 4.93$, $p < 0.01$) but not when incorrect "scrambled" responses were considered (2.25 vs. 2.28; $t(17) < 1$). No difference was observed for ratings made to scrambled images (2.36 vs. 2.37; $t(17) < 1$). In other words, the good/bad distinction was only apparent when participants correctly detected coherent structure in the image.

Experiment 2: Clarity

Experiment 1 documented a significant performance advantage for good category exemplars over bad category exemplars in a non-categorization task. However, participants were provided with a list of the categories used in the experiment during the delivery of the task instructions, and also performed a secondary rating task that required categorization of the scenes. It is possible that these manipulations influenced performance on the intact vs. scrambled detection task, for instance, by encouraging observers to first attempt to categorize each scene, and then make an “intact” response only if they could successfully do so.

This experiment tests whether such instructions and secondary category task were necessary for Experiment 1's results. Specifically, it tests whether the advantage for good exemplars would persist when a different rating task was used that did not evoke image category. To further lessen the likelihood of participants relying on category information, they were not told of the categories used in the experiment, nor were they given any indication that a specific set of categories would be used. Instead of a secondary category judgment, participants were asked to rate the clarity of each image. If participants are less likely to see a bad category exemplar as intact, it is possible that their subjective experience will reflect this. This subjective experience was quantified by replacing the category judgment used in the first experiment with a judgment of clarity.

Method

The design of Experiment 2 was the same as Experiment 1, with two exceptions. Instead of being prompted to rate how well the image exemplified its category,

participants were prompted to provide a rating of how “clearly” they felt they saw the image on the preceding trial by pressing a number between one (not clearly) and five (very clearly) immediately after each intact/scrambled response. Moreover, participants were never given a list of the categories used in the experiment; they were simply told that they would be looking at pictures of scenes.

Results and Discussion

Participants' overall accuracy on the intact/scrambled distinction was 85% and they achieved this accuracy with an average image durations of 45 ms. We again observed a significantly greater d’ for good exemplars than bad exemplars (2.40 vs. 2.12; $t(17) = 3.81$, $p < 0.01$), indicating that the category rating judgment required in Experiment 1 was not necessary for the good exemplar advantage. Indeed, a mixed-design ANOVA with one within-subject factor (good vs. bad category exemplars) and one between-subject factor (Experiment 1 vs. Experiment 2) found no main effect of experiment ($F(1,34) < 1$) and no interaction between experiment and exemplar quality ($F(1,34) < 1$) on participants’ d’. As in Experiment 1, greater sensitivity for good than bad exemplars was apparent for all categories except for forests, which once again showed the lowest overall sensitivity (see Figure 2b).

A good/bad effect was also seen in participants’ clarity ratings. Intact good images were rated as more “clear” than intact bad images (3.62 vs. 3.41; $t(17) = 5.21$, $p < 0.01$; Figure 3b), while no difference was observed in the clarity ratings of scrambled good and bad exemplars (2.40 vs. 2.38; $t(17) < 1$). In other words, not only did good images result in higher sensitivity to the intact vs. scrambled distinction, participants experienced them

as being more clear in keeping with our hypothesis that good exemplars are actually perceived more readily than bad exemplars. Taken together, these results imply that participants tend to see images that are good examples of a basic-level scene category more clearly than bad, regardless of whether they are asked to perform a covert categorization task.

Experiment 3: Inversion

Given the speed at which neural signals can distinguish scene categories, it could be the case that scene perception is mediated by an ultra-fast template matching process. Good exemplars, by this account, may be better distinguished from scrambled images by virtue of them being more similar to a prototypical image of their category than bad exemplars. In other words, it may be that scenes rated as “good” better reflect the environments with which we have more experience, and therefore correspond to stimuli to which the visual system is best adapted. In Experiments 1 and 2, intact images of good category exemplars were easier to discriminate from scrambled images; if the good/bad difference is driven by such a template-matching process, then a manipulation that disrupts an image’s ability to conform to an image-like template would diminish the effect.

One possibility for disrupting an image’s similarity to a pre-existing template is inversion (e.g. reflection around the horizontal axis). Presumably if we have templates for natural scene categories they will contain orientation information relative to our everyday environment; after all, the vast majority of natural scenes we encounter will be upright with respect to our head and body. Thus, inverting an image of a natural scene should affect its ability to conform to any such templates since large regions of the image will now be out of register with the template. Consistent with the idea that we may have ecologically oriented templates of natural scenes, image inversion has also been shown to impair observers’ ability to categorize images of natural scenes (Walther et al., 2009), similar to the inversion effect for faces (Yin, 1969). To test whether our good/bad effect

reflects the degree to which an image matches an ecologically oriented scene template, in Experiment 3, an inversion factor was introduced into the design of the previous experiments. If this form of template matching underlies the good/bad effect, inversion should reduce the difference in performance between good and bad exemplars. Moreover, if the extent to which a scene matches an upright category template really does influence how well it can be discriminated from a phase-scrambled image, then observers should perform more poorly on inverted images than upright images.

Methods

The design of Experiment 3 was similar to that used in Experiment 1 and 2. Since a between-subjects comparison of the previous experiments failed to find an effect of secondary task on scene detection performance, none was used here; trials concluded as soon as participants provided an intact vs. scrambled judgment. As in Experiment 2, the list of categories used in the study was withheld from participants until the end of the session, along with the fact that all images were exemplars from a specific set of categories. Since trials were shorter, participants now performed 20 testing blocks, for a total of 29 blocks in a single session. Staircasing blocks (9 in total) were still comprised of upright “middle” images as in the previous experiments; however, on 50% of the intact-image trials in the testing blocks, the photograph was inverted by reflecting it about its horizontal axis. Because inversion preserves the amplitude spectrum of the natural scenes, it was not necessary to invert the scrambled images, which still accounted for 50% of the total trials.

Results and Discussion

The design of Experiment 3 contained two factors: images could be good or bad exemplars of their category, and they could appear upright or inverted. If the degree to which an image matches an upright template is related to observers' ability to discriminate it from a scrambled image, then inverting scenes should impair performance, producing a main effect of inversion. Additionally, if the good/bad effect found in Experiments 1 and 2 was due to differences in the ability of good and bad images to match such a perceptual template, then inverting scenes should reduce this effect, resulting in an interaction between good/bad and upright/inverted. Note that these possibilities are independent; inversion could impair discrimination performance overall but fail to modulate the difference between good and bad exemplars. Participants' performance is shown in Figure 4. A two-way repeated measures ANOVA was fit to participants' d', revealing a significant main effect for both good vs. bad category exemplars ($F(1,18) = 4.44$, $p < 0.05$), a replication of the previous experiments. Moreover, there was a significant benefit for upright compared to inverted presentation ($F(1, 18) = 21.02$, $p < 0.01$), consistent with the idea that the degree to which an image matches an upright category template influences its detection. However, no interaction between inversion and category representativeness was apparent ($F(1, 18) < 1$), which implies that the intact/scrambled discrimination benefit for good images is not reliant on a straight-forward match with an upright template.

Scene Similarity

Experiments 1 – 3 showed that the "representativeness" of an image predicts how well it will be detected in an intact vs. scrambled judgment task. Given the brief presentations used in these experiments and the rapid speed of natural scene processing, it is likely that any effect of typicality on scene detection takes place at an early stage of visual processing. It may be the case that the human visual system is tuned to "typical" environments so that they can be processed with greater efficiency, resulting in a relationship between the typicality of a scene and the speed and accuracy with which it is processed. On the other hand, it is possible that the mechanism responsible for this effect operates over shorter time-scales; for example, good scenes may be more effectively primed by the targets from preceding trials, resulting in the good-bad effect.

This section considers both of these possibilities by examining the effects of typicality and priming on scene detection accuracy. Although these two factors suggest different mechanisms underlying the good-bad effect reported in Experiments 1-3, the evaluation of both will rely on the measurement of the similarity between pairs of scene images. The typicality of each image was estimated by computing its average similarity to the remaining images in its category. Priming effects were tested by examining the similarity between the target image of each trial and its predecessor, irrespective of category. Therefore, a computational analysis of the similarity between pairs of natural scenes was first performed on the entire set of images for which ratings were collected. Several different methods of computing scene similarity are considered, and one was selected on the basis of its ability to differentiate the six basic-level categories present in

the image set. Then, typicality and priming effects were evaluated in turn using this similarity measure.

Similarity

Four different measures of image similarity were considered: the correlation of pixel (red, green, and blue) values, the correlation of pixel intensities, the structural similarity image metric (SSIM; Wang et al., 2004), and the L2-norm (euclidean distance) between features extracted according to the spatial envelope (SE) model of scene representation (Oliva & Torralba, 2001). Each of these measures was then evaluated by comparing the similarity of same-category image pairs to that of different-category pairs, and the measure that showed the greatest difference in its values for these cases was used for the remainder of the analyses.

Pixel-wise correlation measures have been previously used with similar images to investigate the pattern of errors in scene categorization (Walther et al., 2009). Inter-category similarity (e.g., between beaches and mountains) was compared to the responses made by participants performing a six-alternative forced-choice scene categorization task, as well as the labels chosen by a statistical pattern recognition algorithm operating fMRI data taken from several regions of interest (ROIs). Pixel-wise similarity was found to predict the pattern of responses made by the classifier using voxels selected from area V1, though it did not correspond to responses in other brain regions or those made by participants. Given its correlation with V1, however, it is conceivable that similarity measured this way might predict priming from image to image.

The structural similarity image metric (SSIM) was designed to provide a measure

of the similarity between a distorted image and the “reference image” from which it was derived (Wang et al., 2004). This measure is meant to take into account the structure inherent in natural images, e.g., the strong correlation between neighboring pixel intensities. Although the SSIM was not designed to measure the similarity of different scenes, its ability to better predict subjective measures of image similarity than mean squared error will possibly make it a more useful similarity measure than pixel-wise correlation (Wang et al., 2004).

Work on the “spatial envelope” (SE) model of scene perception has shown that spectral information can be used to both describe spatial properties of scenes, such as “openness” and “naturalness”, and also categorize scenes at the basic level (Oliva & Torralba, 2001). Images were first rescaled to a square aspect ratio, and spectral information was extracted by calculating the response to gabor filters at three spatial frequencies and eight orientations over a fixed window size of 75 × 75 pixels. The filter responses for each window were concatenated to obtain a feature vector for each image, and the euclidean distance between pairs of vectors provided a measure of dissimilarity. This dissimilarity measure was then subtracted from 1 to provide a measure of similarity, giving identical images a score of 1 and less-similar (more distant) pairs lower values.

Each of the above measures was taken at three spatial scales; the images' original 800 × 600 pixel resolution was used, as well as images down-sampled to 400 × 300 and 200 × 150 pixels. Because both the SSIM and SE measures operated over fixed window sizes, this had the effect of increasing the size of the region in each scene over which the measures pooled data. Each of the four similarity measures was evaluated at these spatial

scales, and the combination that performed best was selected for subsequent analyses.

To determine which of the preceding similarity measures might best capture something akin to human similarity judgments we used the following metric: we chose the measure of similarity between category exemplars that had the highest similarity values between images from the same category compared to images from different categories. To evaluate the similarity measures above, 10,000 pairs of images were randomly selected, with replacement, from the original set of 4025 images from which good, bad, and middle subsets were derived. Each measure was normalized before comparing the mean same-category vs. different-category similarity scores. The largest same- vs. different-category similarity difference was observed for the spatial envelope measure at a spatial scale of 400 × 300 pixels. A two-sample t-test showed that this difference between same- and different-category similarity was significant, $t(9998) = 19.6$, $p < 0.01$.

Typicality

Previous work employing the same images used in this research has shown that "good" category exemplars (those having high representativeness scores) are more accurately categorized in a rapid scene categorization task (Torralbo et al., in prep). Given that stimuli that are more typical of their category are easier to categorize than those that are less typical (Rosch, Simpson, & Miller, 1976), this result suggests that the "representativeness" judgments provided by observers were closely related to the typicality of the scenes. However, Experiments 1-3 found that these representativeness scores also predict how well participants can detect briefly presented scenes and

discriminate them from meaningless images. This finding could be interpreted as evidence that a relationship exists between typicality and detection, in which images that are easier to categorize are also the easiest to “see”. On the other hand, it may be the case that the representativeness ratings that were used to create separate groups of “good” and “bad” images were not based on typicality, but instead some other property of the scenes such as clutter or simplicity, and that this property predicts both categorization and detection of briefly presented scenes. Here, an estimate of each scene's typicality is used to assess whether images are more easily detected by observers when they are more typical.

Using the Spatial Envelope measure of image similarity, the “typicality” of each image was calculated as its mean similarity to all same-category images (Nosofsky, 1986). The typicality of intact trial images was then fit to observers' accuracy data using a simple linear regression. Results show that typicality is a reliable predictor of accuracy ($\beta = 0.15$, $p < 0.01$); participants were better at detecting more typical intact images. The data were then fit to a multiple regression with two factors: typicality, and representativeness (“good” vs. “bad”). Both factors were significant predictors of accuracy, (representativeness: $p < 0.05$; typicality: $p < 0.01$) while the interaction term was marginal but failed to reach significance ($p = 0.09$). In other words, participants were more accurate at correctly detecting intact images that were high in typicality as well as those that were “good”. A hierarchical linear regression (HLR) analysis was then used to determine whether the inclusion of this representativeness factor resulted in a model that better explained the data. The model including both typicality and representativeness was

found to provide a significantly better fit to the data, $F(2) = 9.3$, $p < 0.01$.

These results show that an objective estimate of an image's typicality predicts how easily observers will be able to detect it in the intact/scrambled judgment task. The HLR analysis, however, indicated that the inclusion of representativeness allows for a better prediction of observers' accuracy, suggesting that typicality alone does not account for the good/bad effect.

Priming

Prototype models of scene perception (Rosch, 1975) predict that pairs of “good” exemplars from the same category should be more similar to each other than same-category pairs of “bad” images. If an inter-trial priming effect is influencing participants' perception of scenes, such that a scene is easier to detect when it follows a trial with a similar scene, then the tendency for good image pairs to have high similarity could alone be responsible for the good/bad effect observed in these experiments. That is, the apparent advantage for good scenes in Experiments 1-3 may simply be due to the fact that they have been more effectively primed by recent images, and not any difference between how these images would be handled by the visual system when seen in isolation. To test this potential explanation for the good/bad effect, the relationship between participants' accuracy and the similarity of successive trial was assessed.

First, the similarity values of same-category image pairs were analyzed to determine whether good pairs were actually more similar than bad pairs. Similarity values were selected from pairs of images that were presented to participants on successive “intact” trials, and had the same category and representativeness labels.

Across all such pairs, good image pairs had a mean similarity value of 0.49, and bad pairs a value of 0.38; a two-sample t-test revealed that this difference was significant, $t(299) = 7.15$, $p < 0.01$. If similarity-based priming is observed, this result suggests that this priming could be related to the effect of representativeness on accuracy, since good images would be more likely to follow a similar scene than bad exemplars.

A simple linear regression was then fit to the data to measure the effect of inter-trial image similarity on accuracy, using the responses to trials from Experiments 1 and 2. Similarity was not a reliable predictor of accuracy on all trials, however, when this analysis is restricted to intact image trials, there was a significant effect of similarity on accuracy ($\beta = 0.04$, $p < 0.05$). To determine the relationship between similarity and the good/bad effect, intact trials accuracy data were fit to a multiple regression with both factors. Both variables were statistically significant predictors of accuracy (good/bad: $p < 0.01$; similarity: $p < 0.05$), though the interaction term was not significant ($p > 0.5$). The addition of the representativeness factor resulted in a model with higher explanatory power than the model fit to similarity alone, as shown by HLR analysis, $F(2) = 12.3$, $p < 0.01$. In other words, although priming from similar images may result in greater accuracy on the intact vs. scrambled image discrimination task, the fact that additional variance in accuracy can be explained by representativeness suggests that priming cannot completely account for the good/bad effect on accuracy.

Discussion

An objective measure of scene similarity was used to predict behavior in two ways. First, similarity was used to derive an index of typicality, to assess whether the

effects of representativeness seen earlier were related to prototypicality. Additionally, it was shown that a target image appearing on one trial can “prime” a similar image in a subsequent trial, and allow observers to better detect it. Neither of these measures, however, can entirely account for the results of Experiments 1-3, as the inclusion of a representativeness factor in a model always provides additional explanatory power.

Presumably, the image ratings provided by AMT workers should correspond roughly to the subjective typicality of each image. Of course, our measures of typicality and even the similarity on which it is based are just proxies for the real typicality of the images and the similarity measures employed by the visual system. Thus, the fact they do not account entirely for the effect of representativeness may be because there are other important factors contributing to subjects' representativeness ratings, or because these typicality and similarity estimates are falling short. Both of these possibilities should be explored in future work; the former through systematic investigations of potential candidate features, such as complexity, and the latter through the use of a much larger and more representative set of images from which typicality can be calculated.

Finally, although the spatial envelope model provided a similarity measure that accounted for a number of effects, it must also be remembered that it provides but one estimate of the veridical similarity between pairs of scenes. It is likely the case that subjective ratings were chosen based on a combination of bottom-up cues, which are modeled by the spatial envelope, along with top-down cues. An explicit comparison between the similarity values provided by models and subjective similarity would also potentially allow for a better estimate of image typicality.

Discussion

Using a two-alternative forced-choice discrimination task, we found that intact photographs of bad exemplars of natural scene categories (Experiments 1-3) and inverted scenes (Experiment 3) are more likely than good category exemplars to be mistaken for phase-scrambled images; that is, good exemplars are actually easier to see as coherent images than bad ones. The similar pattern of results observed in the category and clarity experiments (Experiment 1 and 2) shows that the good exemplar advantage is not dependent upon being instructed to consider the category of an image. The effect of scene inversion (Experiment 3) indicates that other factors can impair performance on this discrimination task. Given that both poor category exemplars and inverted scenes have previously been shown to be more difficult to categorize (Walther et al., 2009; Torralbo et al., in prep), our results imply that the extent to which an image makes contact with its category is related to how well it is perceived.

Previous research has documented a relationship between the detection and categorization of an object in a scene (Grill-Spector & Kanwisher, 2005). The detection task employed in this research consisted of a two-interval forced choice task, in which participants determined which of two images were intact and scrambled. Note that, while participants were instructed to detect an object, these stimuli would allow them to base their decisions on which scene appeared intact irrespective of the objects it contained. On each trial, participants also performed a two-alternative forced choice categorization task. When participants were categorizing images as containing cars vs. faces, not only were accuracies and RTs in both tasks comparable, but performance on each task predicted

performance on the other; i.e. correct categorization predicted correct detection, and vice versa. Although recent research has illustrated that categorization performance is worse than detection performance for other object categories (Mack & Palmeri, 2010), here we offer new evidence for a relationship between detection and categorization: images that are difficult to categorize – those that are bad exemplars of their category or inverted – are more difficult to "detect".

What information are observers using to discriminate intact from scrambled images? Recent work has shown that the layout of a natural scene is very easily extracted from briefly presented images (Greene & Oliva, 2009b), making it accessible to observers in the timescales seen here. The inversion effect observed in scene categorization performance (Walther et al., 2009) is likely related to inversion's disruption of both the holistic scene layout and the arrangement of objects within it. Moreover, inversion was also found to make scenes more difficult to detect, suggesting that layout has an important role in the intact/scrambled task as well.

If layout information is driving the difference in good-bad scene detection sensitivity (e.g., good images have more easily-apprehended layout than bad images), then disrupting scene layout through inversion should reduce the effect. However, this was not the case in Experiment 3: main effects of good vs. bad and upright vs. inverted were observed, but there was no interaction between these factors. These data suggest that although layout has a large effect on how well we see the scenes (i.e. there is a large effect of inversion on d'), it plays little role in the difference between good and bad scenes (i.e. no interaction of inversion and good vs. bad). Recent work has shown that a

lack of interaction between inversion and representativeness in a scene categorization task (Caddigan et al., in prep), indicating that the good/bad effect must be driven by some other aspect of scenes that is comparably important to the visual system, but invariant to inversion. Completely unlocalized features are sufficient for computational classification of natural scenes (Li & Perona, 2005; Oliva & Torralba, 2001), although human categorization performance with such information is poor (Loschky et al., 2007, Loschky & Larson, 2009). Moreover, the similarity analysis used here suggests that some degree of localization is necessary to capture the differences between categories used in these experiments.

Previous work has investigated the patterns of neural activation associated with scene perception, evaluating in several regions of interest both the ability to "decode" category from their activity and the similarity of the decoded information to human categorization performance (Walther et al., 2009). Above-chance decoding was observed in a number of areas, including primary visual cortex and the parahippocampal place area (PPA; Epstein & Kanwisher, 1998), with the latter having both the highest decoding rates and the best match with behavior. More recently, it has been shown that in these same areas, the patterns of activity evoked by good category exemplars is decoded more accurately than that elicited by bad exemplars (Torralbo et al., in prep), implying that good exemplars result in a more robust neural representation of scene category. Such a representation may rely on good category exemplars containing some information that allows them to cohere more quickly, which would also explain why good category exemplars are better detected and categorized. Given the speed with which natural

images can be distinguished by the brain (Thorpe et al., 1996) and the brief presentations used here (in some cases, less than 20 ms), the advantage for good exemplars observed in these experiments must be conferred very rapidly, perhaps during the feed-forward sweep of visual information through the visual processing hierarchy (VanRullen & Thorpe, 2001) or by very fast back-projections from frontal areas (Bar, 2003).

Analysis of the images used in Experiments 1-3 revealed two interesting effects. First, performance on a given trial is influenced by the immediately preceding trial, such that an intact image is more likely to be detected when it follows a visually similar image. This positive priming effect for masked natural scene images has not been documented in the literature, and warrants further investigation. However, results from a hierarchical linear regression suggest that this effect is not responsible for the detection advantage enjoyed by good natural scene category exemplars.

More closely related to the effect of representativeness, the pairwise similarity of images can also be used to derive a measure of each image's “typicality”, which also predicts performance on the intact vs. scrambled scene detection task. That is, if the typicality of an image is taken to be its average similarity to other images drawn from the same category, images with high typicality are easier to detect than those with low typicality. Although this effect also fails to account for all the variance described by the good vs. bad factor derived from ratings provided by a separate group of observers, it is likely that subjective typicality is determined using one's lifetime of experience with the natural world and images captured in it, and the set of images used in this research failed to span the space of possible scenes.

However, since typicality values are derived from similarity, this “typicality” result may be an indication of another effect. In contrast with a true relationship between an image's category and how well it is “seen”, it could be the case that perceptual learning is occurring during the experiment, with observers forming an association between some set of features and correct “intact” responses without actually “seeing” the scenes. There are a number of reasons to believe that this might not be the case. First, subjective ratings of clarity collected in Experiment 2 suggest that good images really are being represented as coherent scenes more often than bad images. Second, no feedback was provided during the testing phase of the experiment, which should act to greatly reduce the amount of learning taking place. Finally, a perceptual learning account of detection task performance would fail to predict the robust inversion effect observed in Experiment 3; inverted images accounted for half of the intact images presented to participants in this experiment, yet observers were significantly less likely to detect these images.

This research has shown that good exemplars of natural scene categories are “seen” better than bad category exemplars. Moreover, inverting images also makes them more difficult to detect without influencing the difference between good and bad images, suggesting that a simple perceptual template does not underlie this effect. Performance was also shown to be influenced by an objective measure of each scene's typicality, which suggests that it may be possible to eventually understand what makes some scenes “good”, and why these scenes are easier to detect. Taken together, these results suggest

that the apprehension of visual information is more closely tied to category membership than previously believed.

Figures

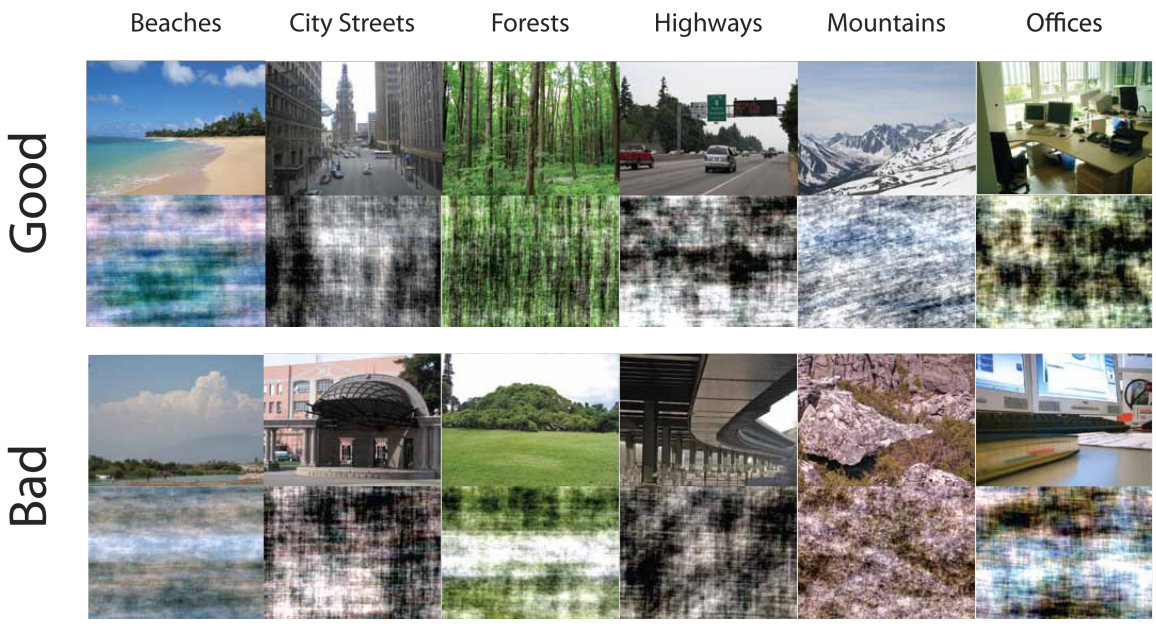

Figure 1: Examples of the stimuli used in the experiments. The top rows show intact and phase-scrambled versions of good exemplars (rated high in category representativeness), the bottom rows contains bad exemplars (with low representativeness ratings). Participants were asked to indicate whether a briefly presented scene was intact or scrambled, irrespective of its category or representativeness.

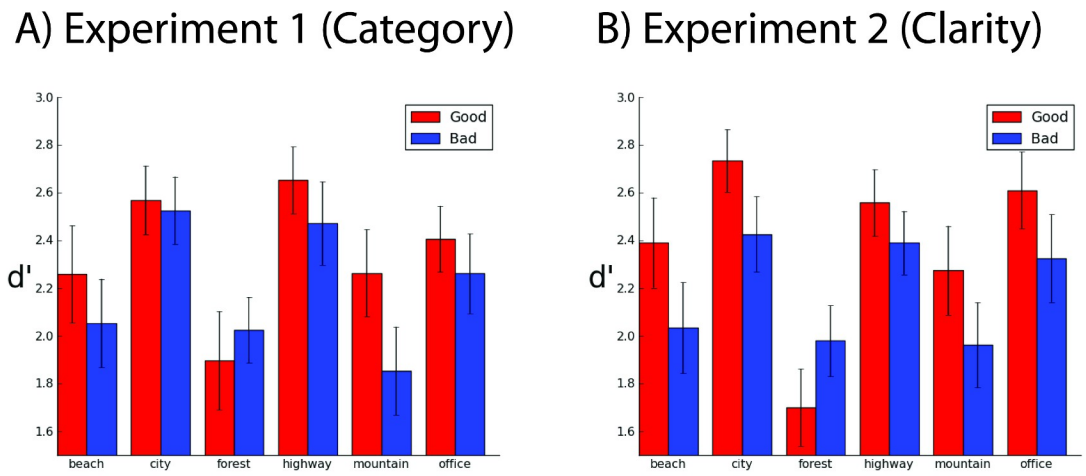

Figure 2: Sensitivity for intact vs. scrambled image discrimination for images that have been rated as high (red, “good” exemplars) and low (blue, “bad” exemplars) in representativeness for their categories. A: d' for each category from Experiment 1, in which participants were given a list of the categories used in the experiment and asked to make a representativeness judgment after each trial. B: d' for each category in Experiment 2, the participants in which were not given a list of categories used in the study, and who were asked to make a judgment of subjective clarity after each trial. Error bars reflect standard error of the mean (SEM).

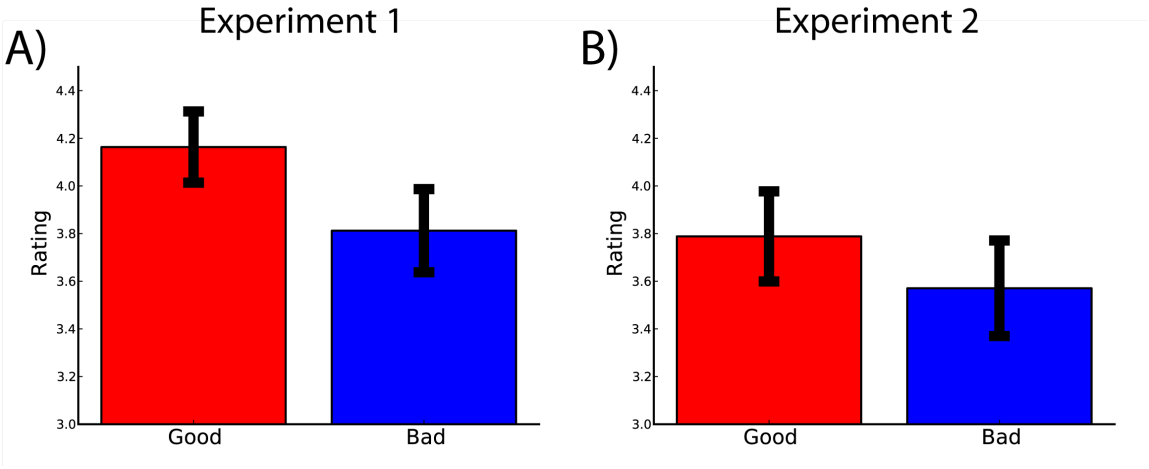

Figure 3: Mean values for the ratings made for good (red) and bad (blue) intact natural scene category exemplars in a secondary rating task. A: Responses to the prompt, “Please enter a number between 1 and 5 to indicate how good an example of its category the photograph was” (Experiment 1). B: Responses to the prompt, “Please enter a number between 1 and 5 to indicate how clearly you saw the photograph” (Experiment 2).

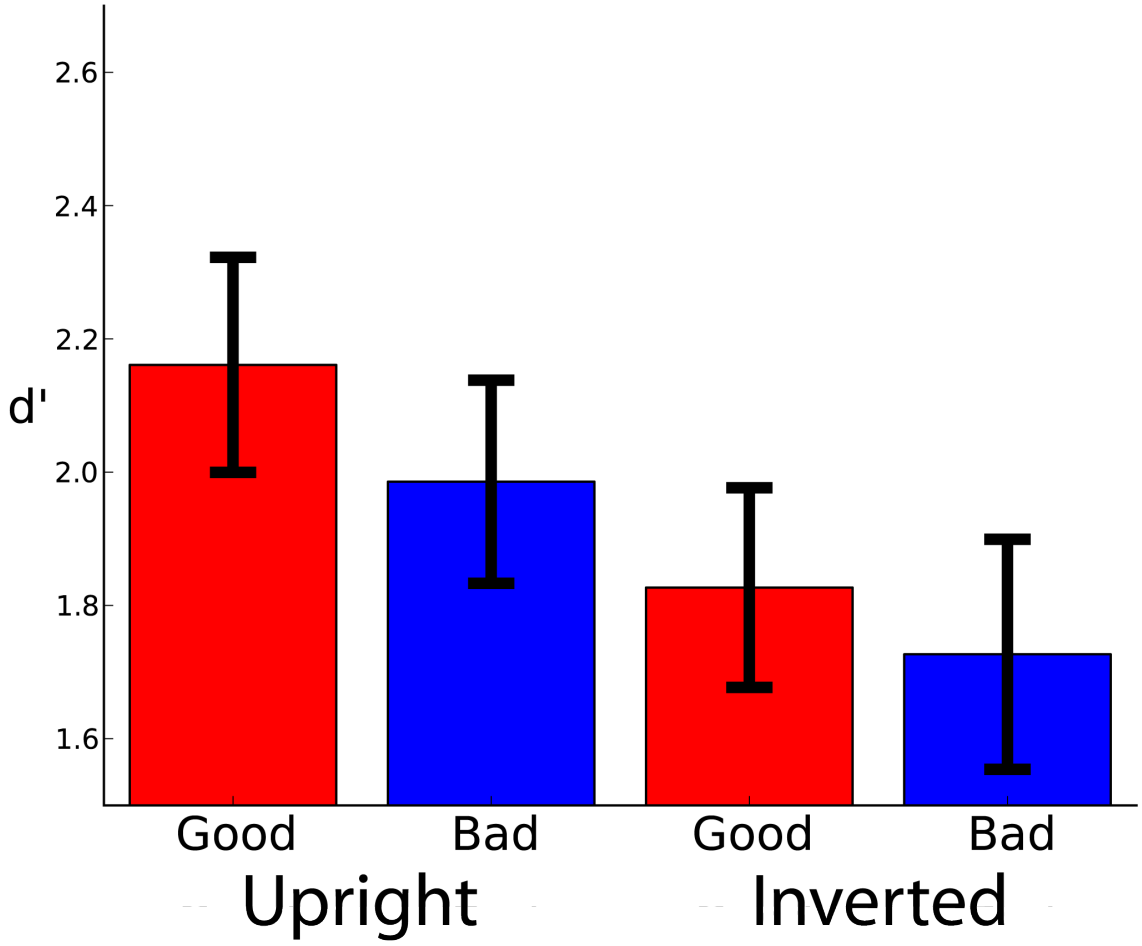

Figure 4: Sensitivity for intact vs. scrambled image discrimination for inverted and upright natural scene images that were rated high (red) and low (blue) in representativeness for their categories.

Works Cited

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience, 15*(4), 600-609.

Blakemore, C., & Tobin, E. A., (1972). Lateral inhibition between orientation detectors in the cat's visual cortex. *Experimental Brain Research, 15*, 439–44.

Caddigan, E., Sahay, T., & Beck, D. M. (in preparation). Natural scenes are robust to bubbling.

Camprodon, J. A., Zohary, E., Brodbeck, V., & Pascual-Leone, A. (2010). Two phases of V1 activity for visual recognition of natural images. *Journal of Cognitive Neuroscience, 22*(6), 1262-1269.

Cutting, J. E., & Vishton, P. M. (1995). Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. (W. Epstein & S. Rogers, Eds.) *Perception of Space and Motion*, 5, 1-37.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*(1), 193-222.

De Valois, R. L., Albrecht, D. G., & Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research, 22*(5), 545-559.

Evans, K., & Treisman, A. (2005). Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance*, 31, 1476–1492.

Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: Recognition, navigation, or encoding? *Neuron*, 23, 115–125.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598-601.

Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. F. (2007). Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience, 19*(9), 1488-1497.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A Optics and Image Science, 4*(12), 2379-2394.

Greene, M. R., & Oliva, A. (2009a). Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cognitive Psychology, 58*(2), 137-176.

Greene, M. R., & Oliva, A. (2009b). The briefest of glances: the time course of natural scene understanding. *Psychological Science*, *20*(4), 464-472.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: as soon as you know it is there, you know what it is. *Psychological Science*, *16*(2), 152-160.

Hancock, P. J. B., Baddeley, R. J., & Smith, L. S. (1992). The principal components of natural images. *Network Computation in Neural Systems*, *3*(1), 61-70.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology, 195*(1), 215-243.

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology, 58*(6), 1233-1258.

Kastner, S., Demmer, I., & Ziemann, U. (1998). Transient visual field defects induced by transcranial magnetic stimulation over human occipital pole. *Experimental Brain Research, 118*(1), 19-26.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye-movements: visual processing speed revisited. *Vision Research*, *46*, 1762-1776.

Kveraga, K., Boshyan, J., & Bar, M. (2007). Magnocellular projections as the trigger of top-down facilitation in recognition. *Journal of Neuroscience, 27*(48), 13232-13240.

Laming, D. (1968). *Information theory of choice-reaction times*. Oxford England Academic Press.

Li, F. F., & Perona, P. (2005). A Bayesian Hierarchical Model for Learning Natural Scene Categories. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR05*, 2, 524-531.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America, 99*(14), 9596-9601.

Loschky, L.C., & Larson, A.M. (2010). The natural/man-made distinction is made prior to basic-level distinctions in scene gist processing. *Visual Cognition, 18*(4), 513-536.

Loschky, L. C., Sethi, A., Simons, D. J., Pydimarri, T. N., Ochs, D., & Corbeille, J. L. (2007). The importance of information localization in scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance, 33*(6), 1431-1450.

Mack, M. L., & Palmeri, T. J. (2010). Decoupling object detection and categorization. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1067-1079.

Maunsell, J. H., Nealey, T. A., & DePriest, D. D. (1990). Magnocellular and parvocellular contributions to responses in the middle temporal visual area (MT) of the macaque monkey. *Journal of Neuroscience, 10*(10), 3323-3334.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39-57.

Oliva, A., & Schyns, P. G. (1997). Coarse Blobs or Fine Edges? Evidence That Information Diagnosticity Changes the Perception of Complex Visual Stimuli. *Cognitive Psychology, 107*(1), 72-107.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42, 145-175.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607-609.

Potter, M. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, 2, 509–522.

Ro, T., Breitmeyer, B., Burton, P., Singhal, N. S., & Lane, D. (2003). Feedback Contributions to Visual Awareness in Human Occipital Cortex. *Current Biology, 11*, 1038-1041.

Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of Experimental Psychology*, 104(3), 192-233.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 491-502.

Rousselet, G., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get to the 'gist' of real-world natural scenes? *Visual Cognition*, 12, 852–877.

Sadr, J., & Sinha, P. (2004). Object recognition and Random Image Structure Evolution. *Cognitive Science, 28*(2), 259-287.

Sanocki, T. (2003). Representation and perception of scenic layout. *Cognitive Psychology*, 47, 43–86.

Schyns, P., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science,* 5, 195–200.

Thorpe, S. J., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381, 520-522.

Torralbo, A., Walther, D. B., Chai, B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (in preparation). Decoding good and bad examples of natural scene categories.

Tversky, B., & Hemenway, K. (1983). Categories of scenes. *Cognitive Psychology, 15*, 121-149.

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience*, *13*(4), 454-461.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of Neuroscience, 29*(34), 10573-10581.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600-612.

Watson, A. B., & Pelli, D. G. (1983). QUEST: a Bayesian adaptive psychometric method. *Perception and Psychophysics*, *33*(2), 113-120.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*(1), 141-145.